
MASSACHUSETTS
LIBRARIES

On Base-Stock Policies for Make-to-Order/Make-to-Stock Production

Viên Nguyen

#3812-95-MSA April 1995

On Base-Stock Policies for Make-to-Order/Make-to-Stock Production

Viên Nguyen
Sloan School of Management, M.I.T., Cambridge, MA 02139

Abstract

This paper analyzes the problem of setting base-stock levels in a production system that operates under both make-to-order and make-to-stock regimes. We study the problem of choosing the base stock level to satisfy a service level constraint and we compare three static priority schemes — first-in-first-out service, priority service for make-to-order jobs, and priority service for make-to-stock jobs. Modeling this system by a mixed queueing network and applying approximations derived from heavy traffic limit theorems, we present an algorithm for setting base-stock levels in each of the three service disciplines and explore conditions under which each service discipline is favorable.

Contents:

April 1995

1 Introduction

The setting of this paper is a production-inventory system that produces multiple types of items. Some product types are produced "to order"; that is, production of these items are initiated by actual customer orders. Other products are made "to stock"; in this case, the system holds finished-goods-inventories from which customer orders are filled, and production of these items is dictated by a policy for keeping the inventories adequately replenished. The mixture of the two types of production, make-to-order (MTO) and make-to-stock (MTS), within a single facility is representative of many manufacturing environments, but the inevitable compromises between capacity investments, inventory costs, and customer service of such an operation are not yet well understood. This paper explores two fundamental issues: how to choose the base-stock levels for make-to-stock items so as to achieve desired service levels, and whether static priority service policies can significantly improve the performance of the system in terms of reducing holding costs while maintaining satisfactory service.

The stochastic nature of the production environment compounded with limited, finite production capacity severely challenges the performance of a production-inventory system. Queueing network models, which naturally incorporate these two characteristics, have been employed successfully in the study of production-inventory systems [3, 17]. Past studies have investigated systems with a large variety of characteristics, but to the author's knowledge, little has been established regarding the analysis of *mixed* production systems [4, 13, 14]. Building on the analysis of a queueing network model of the mixed production system, this paper presents guidelines for setting base-stock levels as well as priority schemes in such a setting.

We assume that the production-inventory system operates as follows. First, production of each type of make-to-stock items follows the "one-for-one replenishment" policy. Under this policy, a base-stock level is specified for each product type; demands are filled from finished-goods-inventory; and each item pulled from inventory triggers a replenishment order to restore the finished-goods-inventory to the desired base-stock level. Second, demands that cannot be met due to insufficient inventory are considered to be lost. We will investigate the performance of the system under three types of static priority service disciplines: first-in-first-out service, priority service for MTO products, and priority service for MTS products.

The paper is organized as follows. Section 2 describes the production-inventory system under study and the corresponding mixed queueing network model. In Section 3 we present formulas for setting base-stock levels. In this section we also investigate the trade-offs between performance measures such as inventory level and lead time for each of the three priority schemes. To simplify the presentation, we consider only production systems with one make-

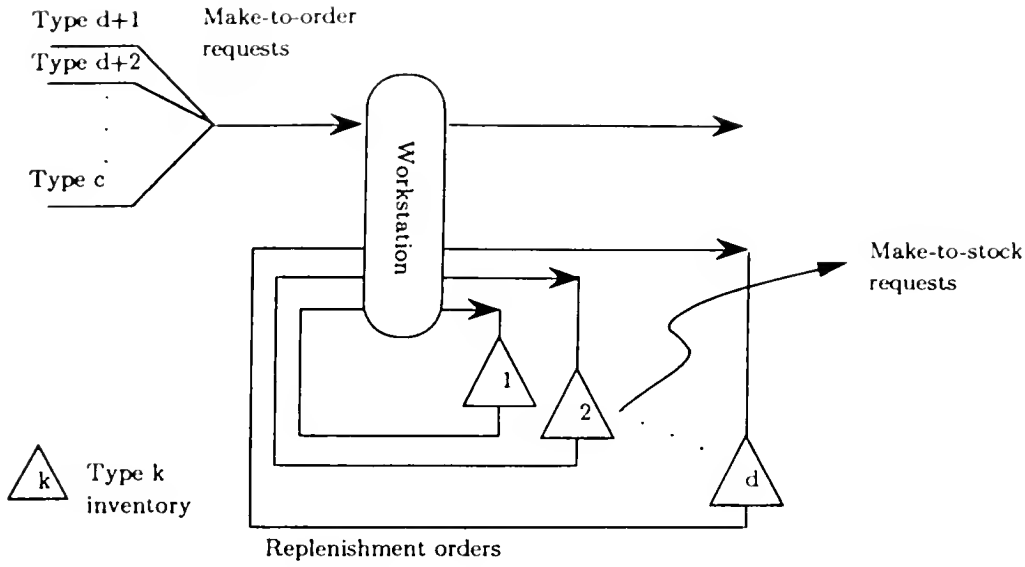

Figure 1: A mixed production-inventory system with multiple job types

to-stock job type in this section. The problem of setting base-stock levels in a production system with multiple make-to-stock job types involves additional technical complications and we return to this problem in Section 5 after we present the theoretical underpinnings of our approximations in Section 4.

2 The Mixed Network Model

The production-inventory system under study is depicted in Figure 1. We envision the production process as a single operation; one can think of the single operation as an aggregation of the entire production process; alternatively, such a characterization would be appropriate if the production process is consistently limited by a single bottleneck workstation. The workstation produces multiple types of products; make-to-stock products are labeled as types $1, \ldots, d$ and make-to-order products are given labels $d+1, \ldots, c$, so that c is the total number of product types in the system. As stated in Section 1, a separate finished-goods-inventory is held for each product type; each make-to-stock item follows the "one-for-one replenishment" policy; and demands that occur when the inventory is empty are considered to be lost.

We model the make-to-order/make-to-stock production system of Figure 1 by the mixed queueing network model with $d+1$ stations depicted in Figure 2. Each station consists of a single server. Station 0 represents the workstation: an arrival at station 0 signals a production request and each service completion at station 0 corresponds to the production of an item.

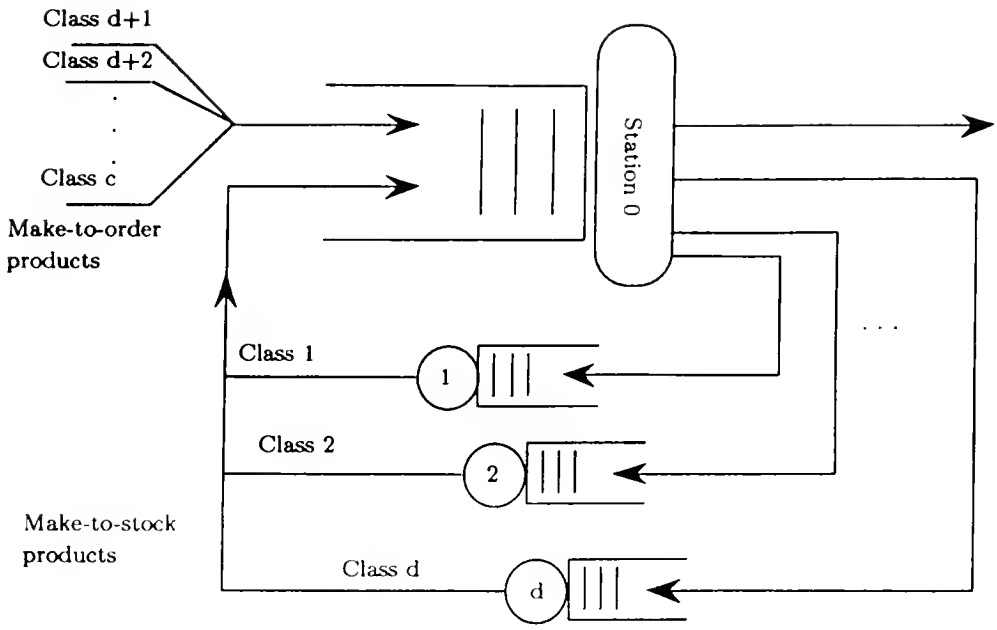

Figure 2: A multiclass mixed queueing network

Stations 1 to d model the finished-goods-inventories (FGI) for make-to-stock products: items in queue k represent FGI of type k ($1 \leq k \leq d$) and service durations at station k correspond to intervals between demands (i.e., inter-demand times) of product k. Each filled demand triggers a corresponding replenishment order, so jobs that "depart" from station k are routed to station 0. Because demands that cannot be filled from inventory are simply lost, the number of items in FGI remains nonnegative; moreover, the number of items in FGI summed with the number of replenishment orders at the workstation for each product type is *constant* at all times and equals the pre-specified base stock level. In the language of queueing networks, make-to-order products are "open" jobs whereas make-to-stock products are "closed" jobs.

Demands for type k products occur at rate λ_k with squared coefficient of variation (SCV) c_k^2; their processing times have mean m_{0k} and SCV c_{0k}^2. We do not impose any restrictions on the distribution of inter-demand times and processing times; moreover, each product type may have its own demand pattern and processing time distribution. For concreteness, we will say that inter-demand times and processing times for each job type form independent sequences of independent and identically distributed random variables. This assumption is made only for ease of exposition; as our results are based on heavy-traffic limits of the network, they are valid under weaker conditions as well (see, for example, [7, 9, 10]). Lastly, we will assume that no setup times are incurred when switching between job types.

Figure 2 suggests that we model the demand processes of make-to-stock products by the

service processes at stations 1 to d. Strictly speaking, a service process characterized by independent and identically distributed service times may not be a faithful representation of the demand process because the first inter-demand time following a period of inventory depletion typically is not statistically similar to other inter-demand intervals. That is, the first service time of a busy period should be characterized by an "excess life" distribution, and this is the same as the inter-demand time distribution if and only if the demand process is Poisson. Nonetheless, the difference is not significant in the sense that the two systems closely approximate each other under heavy traffic conditions (see [11, 13, 14]).

We now introduce notation that will be used in the remainder of this paper. Write n_k, $k = 1, \ldots, d$, to mean the the target base-stock level of type k make-to-stock products. We will assume that initially, the finished-goods inventory of each make-to-stock product type is full. The relative traffic intensity at the workstation is given by

$$\rho_0 \equiv \sum_{k=1}^{c} \lambda_k m_{0k}. \tag{1}$$

Let us write ρ_s (respectively, ρ_r) to mean the contribution from make-to-stock (respectively, make-to-order) products to the relative traffic intensity at the workstation:

$$\rho_s \equiv \sum_{k=1}^{d} \lambda_k m_{0k} \qquad \rho_r \equiv \sum_{k=d+1}^{c} \lambda_k m_{0k}. \tag{2}$$

We are interested in networks that are neither pure open nor pure closed networks, which implies $\rho_s > 0$ and $\rho_r > 0$. Next, define

$$n \equiv \sum_{k=1}^{d} n_k \quad \text{and} \quad \beta_k \equiv \frac{n_k}{n}. \tag{3}$$

The fill rate of type k jobs, defined to be the fraction of type k orders that are filled from inventory, is denoted by α_k. The number of type k jobs in the workstation at time t is denoted by $Q_{0k}(t)$, which one can interpret as type k work-in-process (WIP) inventory. The number of items in type k finished-goods inventory, $k = 1, \ldots, d$, is denoted by $Q_k(t)$ and satisfies the relationship $Q_k(t) = n_k - Q_{0k}(t)$. Finally, $W(t)$ is the amount of work present at the workstation at time t, which includes the remaining processing time of any job in service at that time.

Product-Form Solutions

It is possible to derive explicit product-form steady-state distributions for this network under the assumptions that

(a) each station operates under the FIFO service discipline;

(b) all processing and inter-demand times are exponentially distributed; and

(c) all product types have the same mean processing time (at the workstation).

(See Kelly [12].) Under these conditions, it is known that the condition

$$\rho_r < 1 \tag{4}$$

is necessary and sufficient for stability, where ρ_r is the contribution from make-to-order jobs to the relative traffic intensity at the workstation. Throughout this paper we will assume that condition (4) holds. This of course implies $\lambda_k < m_{0k}^{-1}$ for each make-to-order product, and we will also assume the same holds for make-to-stock products. Moreover, we will assume that make-to-stock products do not require full utilization of production capacity — namely, $\rho_s < 1$ as well — although this is not strictly required for stability.

3 Setting the Base-Stock Levels

Let us begin by addressing the problem of setting base-stock levels in a system with one make-to-stock job type ($d = 1$), deferring the general discussion of multiple make-to-stock job types to Section 5. (The number of make-to-order job types is not restricted, so $c \geq 2$.) Although the formulas become more complicated with multiple make-to-stock job types, the behavior of the system is qualitatively similar.

Denote by α_1 the desired fill rate for MTS jobs. The problem is to determine $N_1(\alpha_1)$, $N_1^r(\alpha_1)$ and $N_1^s(\alpha_1)$ the minimum base-stock level required to achieve this service level under FIFO service, priority service for make-to-order jobs and priority service for make-to-stock jobs, respectively.

Base-Stock Approximations

Define

$$\sigma^2 \equiv \sum_{k=1}^{c} \lambda_k m_{0k}^2 (c_k^2 + c_{0k}^2), \tag{5}$$

with the interpretation that σ^2 is a measure of the variability of the workcenter's *aggregate* incoming workload. Similarly, define

$$\sigma_s^2 \equiv \sum_{k=1}^{d} \lambda_k m_{0k}^2 (c_k^2 + c_{0k}^2), \tag{6}$$

interpreting σ_s^2 as a measure of the variability of the workcenter's incoming *make-to-stock* workload. Let $\lfloor x \rfloor$ denote the integer part of a real number x. We propose the following approximations for the base-stock levels:

Approximation 1 (Base-stock Levels)

$$N_1(\alpha_1) = \begin{cases} \left\lfloor \frac{\lambda_1 \sigma^2}{2\lambda_1 m_{01}} \left(\frac{\alpha_1}{1-\alpha_1} \right) \right\rfloor + 1 & \text{if } \rho_0 = 1 \\ \left\lfloor \left(\frac{\lambda_1 \sigma^2}{2\rho_0(1-\rho_0)} \right) \alpha_1 \ \ln \left(1 + \frac{1-\rho_0}{\lambda_1 m_{01}(1-\alpha_1)} \right) \right\rfloor + 1 & \text{otherwise;} \end{cases} \tag{7}$$

$$N_1^r(\alpha_1) = \left\lfloor \frac{1}{1-\rho_r} N_1(\alpha_1) \right\rfloor + 1; \tag{8}$$

$$N_1^s(\alpha_1) = \left\lfloor \left(\frac{\lambda_1 \sigma_s^2}{2\rho_s(1-\rho_s)} \right) \alpha_1 \ \ln \left(1 + \frac{1-\rho_s}{\lambda_1 m_{01}(1-\alpha_1)} \right) \right\rfloor + 1. \tag{9}$$

The formulas given in (7)–(9) are based on heavy-traffic analysis of the queueing network model. This analysis, which we present in next section, gives *exact* characterizations of the system's dynamics when the workstation is balanced ($\rho_0 = 1$) and the aggregate base-stock level $n \to \infty$, a state of affairs that we refer to as the "heavy traffic limit." We then modify the exact formulas to obtain approximations for systems not in the heavy traffic limit with $n < \infty$ and ρ_0 not necessarily equal to one. The accuracy of these heavy-traffic based approximations were investigated in Nguyen [13, 14], which showed that for FIFO systems, the method provides good approximations for performance measures such as throughput rates, inventory levels and lead times. The next section of this paper will present further simulation experiments, including networks with priority service, that also testify to the accuracy of the approximation method.

Encouraged by these preliminary experiments, we now turn our attention to the qualitative implications of equations (7)–(9). We are primarily interested in the relationships between base-stock levels, service levels, inventory levels and priority schemes.

Let us first consider equation (7), the base-stock level under FIFO service. First, the base-stock level is approximately proportional to the fill-rate function $\frac{\alpha_1}{1-\alpha_1}$, which is directly evident in the case of $\rho_0 = 1$ and can also be verified for $\rho_0 \neq 1$ by expanding the log term. Second, a higher base-stock level is required for a system with higher aggregate variability, as measured by the unitless term

$$\lambda_1 \sigma^2 = \sum_{k=1}^{c} \lambda_1 \lambda_k m_{0k}^2 (c_{ak}^2 + c_k^2).$$

Third, the base-stock level increases when type 1 utilization, as measured by $\lambda_1 m_{01}$, decreases, or equivalently, because ρ_0 is approximately one, when there is more competing work.

Taking FIFO as the "standard" scenario, let us now compare the base-stock levels under the two priority disciplines. Our approximation method reveals a very simple relationship between FIFO and MTO priority: To attain the same service level when priority is given to MTO jobs, set the base-stock level to be $\frac{1}{1-\rho_r}$ times that of FIFO.

As an example, suppose that MTO jobs do not command a high portion of the workcenter's capacity (say $1-\rho_r = 0.90$). If the workstation were to give priority to MTO products, the same MTS service level can be attained with little increase in the base-stock level (the base-stock level would increase by 11%), while MTO lead time would be significantly reduced.

When MTS jobs receive priority, the base-stock level is set as if there were no make-to-order jobs in the system. Hence, the "aggregate" variability σ^2 is replaced by its MTS component σ_s^2, and utilization ρ_0 of the workstation is reduced to consist only of MTS work ρ_s.

Trade-offs Between Throughput and Inventory

To get a better a sense of the trade-offs between throughput and inventory, let us consider a network with one make-to-order and one make-to-stock job type ($d = 1$ and $c = 2$). We will consider three such systems. In all three systems, all inter-demand times and processing times in the network are exponentially distributed; the mean processing time for both job types is 0.0008; and the utilization of the workstation is 0.96. In the first system, we set $\lambda_1 = 1000$ and $\lambda_2 = 200$; in the second system $\lambda_1 = \lambda_2 = 600$; and in the third system, $\lambda_1 = 400$ and $\lambda_2 = 800$. With these parameters, MTS utilization varies from 0.80 in System 1 to 0.50 in System 2 to 0.32 in System 3.

Using the approximations developed in Section 4, we can quickly and effectively estimate the performance of systems with different parameters operating under different service disciplines. For MTS jobs, the primary performance measures of interest are fill rate and average finished goods inventory. For MTO jobs, we are interested in the average lead time, by which we mean the sum of the time that an order spends waiting to be processed and its processing time. Rather than reporting the average lead time, however, we will present our results in terms of the "waiting factor," defined as the ratio of the average waiting time to the average service time, which we feel is a more meaningful characterization of the penalty due to congestion.

Figure 3 shows the average FGI and the average waiting factor for each priority scheme and each system when a fill rate of 0.90 is required for MTS jobs. Each line in the graph corresponds to one of the three systems; the left-most data point of each line corresponds to the policy that gives priority to MTO jobs; the middle point corresponds to FIFO service; and the right-most point arises from the policy that gives priority to MTS jobs. Figure 4 shows

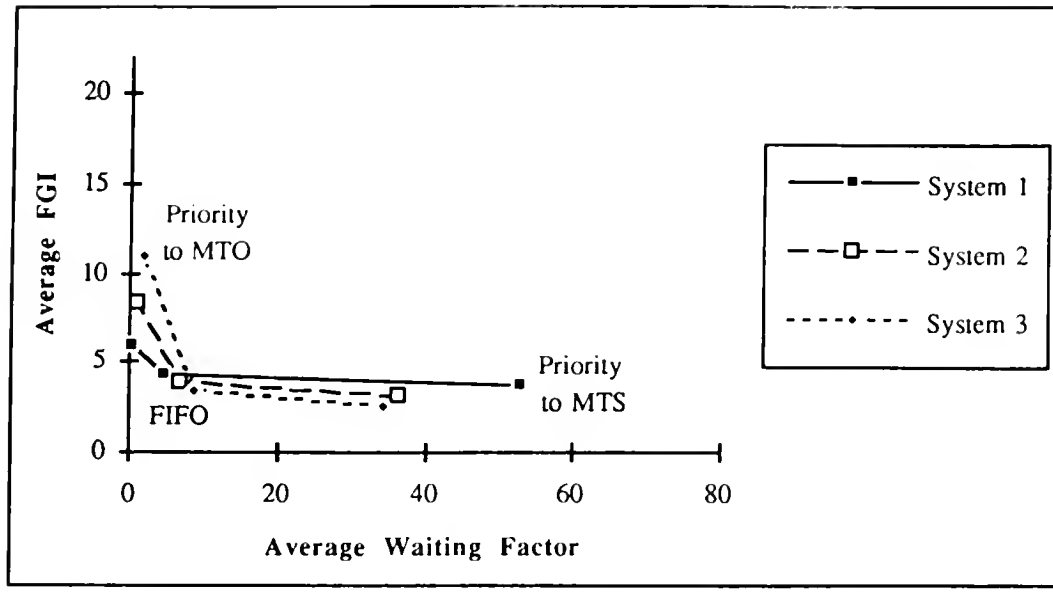

Figure 3: Attaining a fill rate of 0.90

the analogous performance measures associated with a fill rate of 0.95.

Let us investigate Figure 4 in more detail, starting with the FIFO data points. System 3 has the lowest percentage of MTS work so we expect it to require the lowest base-stock for a given service level constraint. As seen in Figure 4, the average FGI is lowest for System 3 and highest for System 1, which has the largest percentage of MTS work. On the other hand, the waiting penalty for MTO jobs is highest in System 3 and lowest in System 1.

The waiting penalty for MTO jobs can be reduced by giving priority to make-to-order jobs, and the MTO data points in Figure 4 show the corresponding increase in average FGI that is required to maintain a fill rate of 0.95. The increase is much more dramatic in System 3 than in System 1, and in fact, under MTO priority, System 3 requires the highest base-stock level. A similar analysis applies to the MTS data points. If it is expensive to hold FGI for MTS jobs, one can decrease the required base-stock level while maintaining the same service level by giving service priority to MTS jobs. This results in a degradation of service for MTO jobs as shown in Figure 4, and the degree of degradation in performance depends on the parameters of the particular system. Depending on the relative costs of holding FGI, costs of having WIP, and costs of customer delay, it may be desirable to operate System 1 by giving priority service to MTO jobs, to operate System 2 with FIFO service, and to operate System 3 by giving priority service to MTS jobs.

To get a better appreciation of the impact of the fill rate constraint, Figure 5 shows the

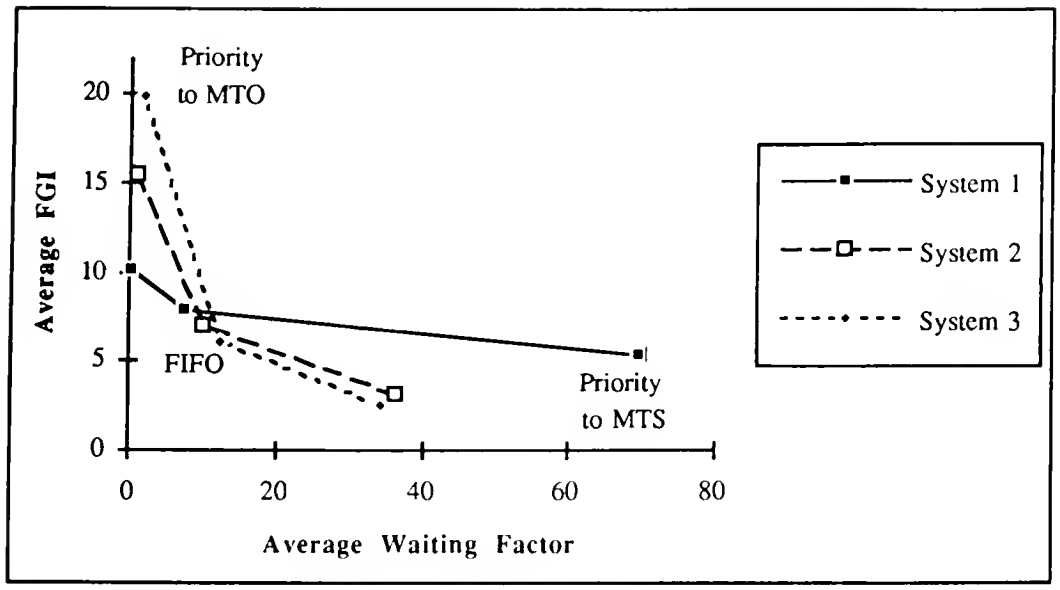

Figure 4: Attaining fill rate = 0.95

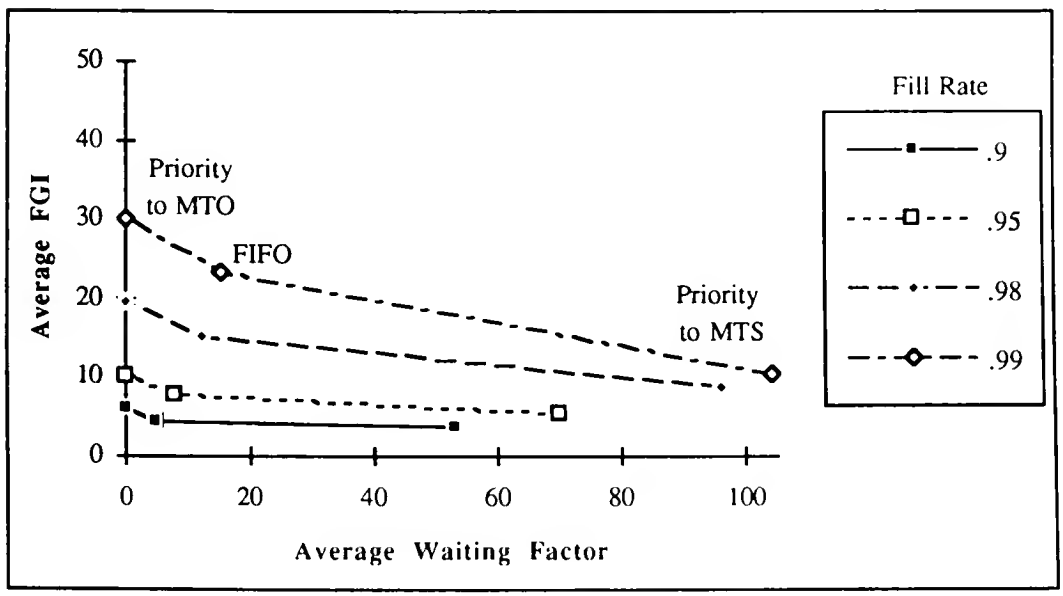

Figure 5: Performance measure trade-offs: System I

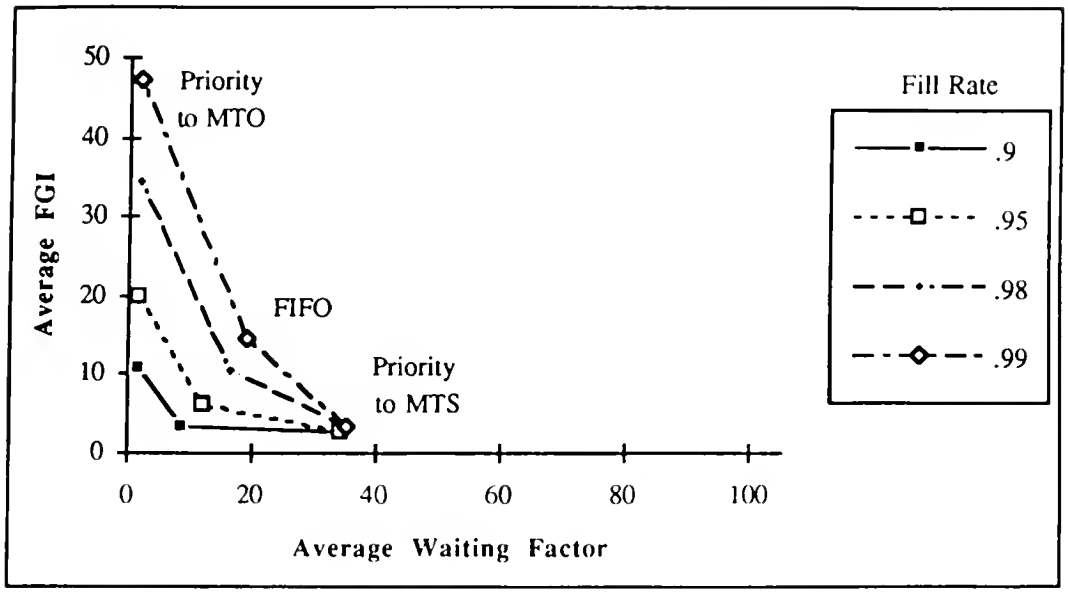

Figure 6: Performance measure trade-offs: System 3

average FGI and average waiting factor in System 1 for varying fill rates, ranging from 0.90 to 0.99. As in the previous graphs, the left most, middle, and right most data points correspond to MTO priority, FIFO, and MTS priority, respectively. Figure 6 shows the analogous results for System 3.

4 Heavy Traffic Limits and Approximations

Technical Preliminaries

In order to state the heavy-traffic limit theorems, we need to refer to a "sequence of systems in heavy traffic." Indexing systems in the sequence by n, we will denote parameters of the n^{th} system by a superscript (n): for example, $\lambda_k^{(n)}$ is the demand rate of type k jobs and $Q_{0k}^{(n)}(t)$ is the type k work-in-process inventory at time t in the n^{th} system. Fixing a vector $(\beta_1, \ldots, \beta_d)$ where $0 < \beta_k < 1$ and $\sum_{k=1}^{d} \beta_k = 1$, we set the base-stock level for type k jobs, $k = 1, \ldots, d$ in the n^{th} system to be $n_k = \beta_k n$. More precisely, we ought to choose an integer value, such as the integer part of $\beta_k n$ plus one. (We add one so that we are guaranteed a positive base-stock level.) In the interest of keeping the exposition simple, however, let us proceed with the former prescription while keeping in mind that this is without any loss representation power.

For the limit theorems, we require that $\lambda_k^{(n)} \to \lambda_k$ and $m_{0k}^{(n)} \to m_{0k}$ as $n \to \infty$. Moreover,

we will assume that the following *heavy traffic condition* holds:

$$n(\rho_0^{(n)} - 1) \rightarrow \mu \text{ as } n \rightarrow \infty \tag{10}$$

where μ is a finite (but not necessarily nonpositive) number. This condition implies that the approximations derived from the limit theorems are particularly relevant when the load at the production facility is approximately balanced and the base-stock levels are large.

The heavy traffic limit theorems will establish convergence of scaled processes as the base-stock level becomes large, i.e., as $n \rightarrow \infty$. Specifically, we are interested in processes that have been scaled in the following manner:

$$\begin{aligned} W^n(t) &\equiv \frac{1}{n} W^{(n)}(n^2 t); \\ Q_{0k}^n(t) &\equiv \frac{1}{n} Q_{0k}^{(n)}(n^2 t), \quad k = 1, \ldots, c; \\ Q_k^n(t) &\equiv \frac{1}{n} Q_k^{(n)}(n^2 t), \quad k = 1, \ldots, d. \end{aligned}$$

The mode of convergence in the limit theorems, denoted by $\Longrightarrow$, is convergence of distributions on the space $\mathbf{D}$, where $\mathbf{D}$ is the space of right continuous functions on $[0, \infty)$ having left limits, and the topology is the Skorohod J_1 topology [2, 16].

First-in-first-out service discipline

We will start by reviewing the case of FIFO service, which was studied in Nguyen [13, 14]. Let us assume that production of all job types follows the first-in-first-out service discipline. Define

$$b^* = \min\{\lambda_1^{-1}\beta_1, \ldots, \lambda^{-1}\beta_d\}.$$

We will write Q_0^n and Q^n to mean the vectors $(Q_{01}^n, \ldots, Q_{0c}^n)$ and $(Q_1^n, \ldots, Q_d^n)$, respectively, and similarly for Q_0^* and Q^*.

Theorem 1 (Nguyen [14], FIFO) *Suppose that the production facility operates under FIFO service and the heavy traffic condition (10) holds, then* $(W^n, Q_0^n, Q^n) \Longrightarrow (W^*, Q_0^*, Q^*)$ *as* $n \rightarrow \infty$, *where*

$$\begin{aligned} &W^* \text{ is reflected Brownian motion on } [0, b^*] \text{ with drift } \mu \text{ and variance } \sigma^2, \qquad (11) \\ &Q_{0k}^*(t) = \lambda_k W^*(t), \quad 1 \le k \le c, \\ &Q_j^*(t) = \beta_k - Q_{0j}^*(t), \quad 1 \le j \le d. \end{aligned}$$

Remark: Henceforth we will express statement (11) by the abbreviated notation "$W^* = RBM([0, b^*], \mu, \sigma^2)$."

Theorem 1 suggests that for large base-stock levels, the workload process at the production facility can be approximated via

$$\frac{1}{n} W(n^2 \cdot) \approx RBM([0, b^*], n(\rho_0 - 1), \sigma^2).$$

Upon reversing the scaling, we obtain the "heavy traffic approximation"

$$W(\cdot) \approx RBM([0, b], \rho_0 - 1, \sigma^2), \tag{12}$$

where

$$b = \min\{\lambda_1^{-1} n_1, \ldots, \lambda_d^{-1} n_d\}. \tag{13}$$

The approximation takes the form of a process that lives on the bounded interval $[0, b]$, whereas the actual workload process may assume arbitrarily large values! This phenomenon, which was discussed in [13, 14], may be explained by saying that make-to-stock jobs, whose volume is limited by the base-stock levels, effectively "regulate" the total workload process so as to keep the amount of work from becoming too large *relative* to the base-stock levels.

The one-dimensional reflected Brownian motion (RBM) essentially captures all of the dynamics of the production-inventory system, which itself is c-dimensional where c is the number of job types. From Theorem 1, we obtain the following approximation for type k WIP inventory:

$$Q_{0k}(\cdot) \approx \lambda_k W(\cdot), \quad k = 1, \ldots, c; \tag{14}$$

and for type k FGI, we have

$$Q_k(\cdot) = n_k - Q_{0k}(\cdot) \quad k = 1, \ldots, d.$$

Closer inspection of the above two equations reveals that type k FGI is an RBM on the interval $[n_k - \lambda_k b, n_k]$. If k is such that $b = \lambda_k^{-1} n_k$, then the approximation states that the FGI behaves as a reflected Brownian motion and takes values between 0 and n_k. However, if k is such that $\lambda_k^{-1} n_k > b$, then the approximation suggests that the FGI process is bounded away from zero. In other words, the heavy traffic approximation suggests that type k jobs *never* stocks out!

Under the heavy traffic scaling and the eventual limit, one or more job types will emerge as "bottleneck job types," which are those job types with the smallest values of $\lambda_k^{-1} n_k$. The same "regulating" mechanism that keeps the workload process from becoming too large relative to

the total base-stock levels also serves to protect non-bottleneck job types from ever stocking out. That is, non-bottleneck FGI's are kept from becoming too small relative to the FGI's of bottleneck job types. (See Nguyen [14] for more details.)

To state our approximation, let us suppose that make-to-stock product types are numbered so that

$$\lambda_1^{-1} n_1 < \lambda_2^{-1} n_2 < \cdots < \lambda_d^{-1} n_d \tag{15}$$

(note that $b = \lambda_1^{-1} n_1$); at the end of this section we will discuss the case where some of the inequalitites may be equalities. In addition, define

$$\theta \equiv \frac{2(\rho_0 - 1)}{\sigma^2}. \tag{16}$$

Approximation 2 (FIFO) *Suppose that the production facility processes jobs on a FIFO basis. The long-run average make-to-order fill rates, make-to-stock FGI's, make-to-order WIP inventory levels, and make-to-order lead times are approximated by* $(\tilde{\alpha}, \tilde{Q}, \tilde{Q}_0, \tilde{T})$, *respectively, defined as follows:*

$$\tilde{\alpha}_1 = \begin{cases} 1 - \dfrac{\sigma^2/2m_{01}}{n_1 + \sigma^2/2m_{01}} & \rho_0 = 1 \\ 1 - \dfrac{1}{\lambda_1 m_{01}} \left(\dfrac{\rho_0 - 1}{1 - e^{-n_1 \rho_0 \theta / \tilde{\alpha}_1 \lambda_1}} \right) & \textit{otherwise.} \end{cases} \tag{17}$$

$$\tilde{\alpha}_k = 1 - \frac{1}{\lambda_k m_{0k}} \left(\frac{\rho_0^k - 1}{1 - e^{-n_k \rho_0^k \theta^k / \tilde{\alpha}_k \lambda_k}} \right), \quad (k = 2, \ldots, d), \quad \textit{where} \tag{18}$$

$$\rho_0^k \equiv \sum_{l=1}^{k-1} \tilde{\alpha}_l \lambda_l m_{0l} + \sum_{l=k}^{c} \lambda_l m_{0l} \quad \textit{and} \quad \theta^k \equiv \frac{2(\rho_0^k - 1)}{\sigma^2}. \tag{19}$$

$$\tilde{W} = \begin{cases} \dfrac{1}{2\lambda_1} \left(n_1 + \dfrac{\sigma^2}{2m_{01}} \right) & \rho_0 = 1 \\ \dfrac{n_1 \rho_0}{\tilde{\alpha}_1 \lambda_1 (1 - e^{-n_1 \rho_0 \theta / \tilde{\alpha}_1 \lambda_1})} - \dfrac{1}{\theta} & \textit{otherwise.} \end{cases} \tag{20}$$

$$\tilde{Q}_k = n_k - \frac{\tilde{\alpha}_k \lambda_k}{\rho_0} \tilde{W}, \quad (k = 1, \ldots, d). \tag{21}$$

$$\tilde{Q}_{0k} = \frac{\lambda_k}{\rho_0} \tilde{W}, \quad (k = d+1, \ldots, c). \tag{22}$$

$$\tilde{T}_k = \tilde{W} + m_{0k}, \quad (k = d+1, \ldots, c). \tag{23}$$

Equations (17) and (18) use as data the fill rates $\tilde{\alpha}_k$, which themselves are unkown and must be approximated. In order for the proposed approximation to be consistent, we therefore must show that equations (17) and (18) uniquely identify $\tilde{\alpha}_k$, $k = 1, \ldots, d$, and that these

parameters take values between 0 and 1. This was done in Nguyen [13]. Moreover, it is straightforward to compute the fill rates numerically.

To end this section, we return to the question of how to resolve possible equalities in equation (15). For example, suppose that $\lambda_1^{-1}n_1 = \lambda_2^{-1}n_2 < \lambda_3^{-1}n_3$. In such a case, we propose to approximate type 2 products in the same manner as type 1 products (as if it were a bottleneck type) and then proceed to type 3 products exactly in the same way as outlined.

Priority for make-to-order products

We now consider the case where production gives priority to products that are made-to-order. The following theorem is proved in Appendix A:

Theorem 2 *Suppose that make-to-order products have preemptive-resume priority in the production facility. Under the heavy traffic condition (10),* $(W^n, Q_0^n, Q^n) \Longrightarrow (W^*, Q_0^*, Q^*)$ *as* $n \to \infty$, *where*

$$
\begin{aligned}
&Q_{0k}^*(t) = 0, \qquad k = d+1, \ldots, c,\\
&W^* = RBM([0, \rho_s b^*],\ \mu,\ \sigma^2),\\
&Q_{0k}^*(t) = \frac{\lambda_k}{\rho_s} W^*(t), \quad k = 1 \ldots, d,\\
&Q_j^*(t) = \beta_j - Q_{0j}^*(t), \quad j = 1 \ldots, d.
\end{aligned}
$$

Under the heavy traffic condition (10) which requires the production faciltiy to be balanced and the base-stock level to be large, Theorem 2 states that the (scaled) number of make-to-order products in the network at any time is negligible. Because make-to-order products receive preemptive-resume priority in production, they are not hindered by make-to-stock work and effectively faces a production center in light traffic. Therefore we find that under the heavy traffic scaling, the number of make-to-stock products at station 1 becomes negligibly small in the limit.

For estimation purposes, however, we propose that make-to-order products be approximated using the standard heavy traffic methodology for GI/G/1 queues that correct for traffic intensities less than unity. Let us assume that job are ordered so that (15) holds (with the understanding that equalities would be treated the same way as described for the FIFO case). Set

$$\sigma_r^2 \equiv \sum_{k=d+1}^{c} \lambda_k m_{0k}^2 (c_{0k}^2 + c_k^2). \tag{24}$$

Paralleling the arguments developed for the FIFO system, we arrive at the following approximation:

Approximation 3 (Priority to Make-to-Order) *Suppose that make-to-order products receive priority in production. The long-run average make-to-order fill rates, make-to-stock FGI's, make-to-order WIP inventory levels, and make-to-order lead times are approximated by* $(\tilde{\alpha}^r, \tilde{Q}^r, \tilde{Q}^r_0, \tilde{T}^r)$, *respectively, defined as follows:*

$$\tilde{\alpha}^r_1 = \begin{cases} 1 - \dfrac{\sigma^2/2m_{01}}{(1-\rho_r)n_1 + \sigma^2/2m_{01}} & \rho_0 = 1 \\ 1 - \dfrac{1}{\lambda_1 m_{01}}\left(\dfrac{\rho_0 - 1}{1 - e^{-(1-\rho_r)n_1\rho_0\theta/\tilde{\alpha}^r_1\lambda_1}}\right). & otherwise \end{cases} \tag{25}$$

$$\tilde{\alpha}^r_k = 1 - \frac{1}{\lambda_k m_{0k}}\left(\frac{\rho_0^{rk} - 1}{1 - e^{-(1-\rho_r)n_k\rho_0^{rk}\theta^{rk}/\tilde{\alpha}^r_k\lambda_k}}\right), \quad (k = 2, \ldots, d), \quad where \tag{26}$$

$$\rho_0^{rk} \equiv \sum_{l=1}^{k-1} \tilde{\alpha}^r_l \lambda_l m_{0l} + \sum_{l=k}^{c} \lambda_l m_{0l} \quad and \quad \theta^{rk} = \frac{2(\rho_0^{rk} - 1)}{\sigma^2}. \tag{27}$$

$$\tilde{W}^r = \begin{cases} \dfrac{1}{2\lambda_1(1-\rho_r)}\left((1-\rho_r)n_1 + \dfrac{\sigma^2}{2m_{01}}\right) & \rho_0 = 1 \\ \dfrac{1}{1-\rho_r}\left(\dfrac{(1-\rho_r)\rho_0 n_1}{\tilde{\alpha}^r_1\lambda_1(1 - e^{-(1-\rho_r)n_1\rho_0\theta/\tilde{\alpha}^r_1\lambda_1})} - \dfrac{1}{\theta}\right) & otherwise. \end{cases} \tag{28}$$

$$\tilde{Q}^r_k = n_k - \frac{\tilde{\alpha}^r_k\lambda_k}{\rho_0}\tilde{W}^r, \quad (k = 1, \ldots, d). \tag{29}$$

$$\tilde{Q}^r_{0k} = \frac{\lambda_k}{2\rho_r}\left(\frac{\sigma_r^2}{1-\rho_r}\right), \quad (k = d+1, \ldots, c). \tag{30}$$

$$\tilde{T}^r_k = \frac{\sigma_r^2}{2(1-\rho_r)} + m_{0k}, \quad (k = d+1, \ldots, c). \tag{31}$$

In this approximation procedure, the make-to-order products are analyzed as if there are no other product types in the system. That is, the approximations in (30) and (31) derive simply from heavy traffic analysis of a multiclass queue (see [9, 10, 15]).

The approximations for make-to-stock products are based on the results of Theorem 2, and the methods for modifying the heavy traffic limit results are essentially similar to those of the FIFO analysis. Observe that in giving priority to the make-to-order products, the workload process at the production center becomes smaller (from Theorem 2, the interval is now $[0, \rho_s b]$ rather than $[0, b]$). This is as we expect: by giving priority to make-to-order jobs, make-to-stock jobs are more likely to be stocked-out, the total rate at which replenishment orders arrive at the workstation decreases, hence the workstation experiences less congestion.

The relationship between queue length and workload is also different from the FIFO case. An application of Little's Law to the equation $Q_{0k}(t) \approx \frac{\lambda_k}{\rho_s}W(t)$ suggests that we interpet $W(t)/\rho_s$ as the *waiting time* of a class k job at the workstation. Let us replace ρ_s by $1 - \rho_r$,

noting that the two are equivalent in the case $\rho_0 = 1$. We now have the natural interpretation of $1-\rho_r$ as the amount of time per unit time available to the workstation for low-priority work. That is, if a low-priority job arrives to see $W(t)$ units of work at the workstation, then it can expect to wait, on average $W(t)/(1-\rho_r)$ units of time before receiving service.

Priority for make-to-stock products

Lastly, we consider the system in which make-to-stock products are given priority in production. The following theorem is proved in Appendix B:

Theorem 3 *Suppose that make-to-stock products have preemptive-resume priority in the production facility. Under the heavy traffic condition (10),* $(W^n, Q_0^n, Q^n) \Longrightarrow (W^*, Q_0^*, Q^*)$ *as* $n \to \infty$, *where*

$$
\begin{aligned}
&Q_{0k}^*(t) = 0, \qquad k = 1, \ldots, d, \\
&Q_k^*(t) = \beta_k, \qquad k = 1, \ldots, d, \\
&W^* = RBM([0,\infty),\ \rho_0 - 1,\ \sigma^2), \\
&Q_{0k}^*(t) = \frac{\lambda_k}{\rho_{0o}} W^*(t), \quad k = d+1 \ldots, c.
\end{aligned}
$$

The limit theorem states that there is no accumulation of WIP inventory for make-to-stock products, which receive preemptive resume production priority and therefore experiences production in light traffic. Moreover, because there is essentially no congestion in production, finished-goods inventory is always full in the heavy traffic limit. As in the previous development, however, we propose that for estimation purposes, make-to-stock products be modeled by the standard heavy traffic approximation for a closed generalized Jackson network [5, 6, 13, 14]. In the following approximation, we assume job types are enumerated so that condition (15) holds. Set

$$\theta_s \equiv \frac{2(\rho_s - 1)}{\sigma_s^2}. \tag{32}$$

Approximation 4 (Priority to Make-to-Stock) *Suppose that make-to-stock products receive preemptive-resume priority in production. The long-run average make-to-order fill rates, make-to-stock FGI's, make-to-order WIP inventory levels, and make-to-order lead times are approximated by* $(\tilde{\alpha}^s, \tilde{Q}^s, \tilde{Q}_0^s, \tilde{T}^s)$, *respectively, defined as follows:*

$$\tilde{\alpha}_1^s = 1 - \frac{1}{\lambda_1 m_{01}} \left(\frac{\rho_s - 1}{1 - e^{-n_1 \rho_s \theta_s / \tilde{\alpha}_1^s \lambda_1}} \right). \tag{33}$$

$$\tilde{\alpha}_k^s = 1 - \frac{1}{\lambda_k m_{0k}} \left(\frac{\rho_s^k - 1}{1 - e^{-n_k \rho_s^k \theta_s^k / \tilde{\alpha}_k^s \lambda_k}} \right), \quad (k = 2, \ldots, d), \quad \textit{where} \tag{34}$$

System	λ_1	λ_2	λ_3	$m_{01} = m_{02} = m_{03}$	ρ_0
I	8.0	6.25	2.0	0.06	0.975
II	10.0	10.0	10.0	0.033	0.990

Table 1: System Parameters: Systems I and II

$$\rho_s^k \equiv \sum_{l=1}^{k-1} \tilde{\alpha}_l^s \lambda_l m_{0l} + \sum_{l=k}^{d} \lambda_l m_{0l} \quad \textit{and} \quad \theta_s^k \equiv \frac{2(\rho_s^k - 1)}{\sigma_s^2}. \tag{35}$$

$$\tilde{W}^s = \frac{n_1 \rho_s}{\tilde{\alpha}_1^s \lambda_1 (1 - e^{-n_1 \rho_s \theta_s / \tilde{\alpha}_1^s \lambda_1})} - \frac{1}{\theta_s}. \tag{36}$$

$$\tilde{Q}_k^s = n_k - \frac{\tilde{\alpha}_k^s \lambda_k}{\rho_0} \tilde{W}^s, \quad (k = 1, \ldots, d). \tag{37}$$

$$\tilde{Q}_{0k}^s = \frac{\lambda_k}{2\tilde{\rho}_0(1 - \rho_s)} \left(\frac{\sigma^2}{1 - \tilde{\rho}_0} \right), \quad (k = d+1, \ldots, c), \tag{38}$$

$$\tilde{T}_k^s = \frac{\sigma^2}{2(1 - \rho_s)(1 - \tilde{\rho}_0)} + m_{0k}, \quad (k = d+1, \ldots, c), \quad \textit{where} \tag{39}$$

$$\tilde{\rho}_0 \equiv \sum_{l=1}^{d} \tilde{\alpha}_l^s \lambda_l m_{0l} + \sum_{l=d+1}^{c} \lambda_l m_{0l}. \tag{40}$$

Note that this approximation of make-to-stock products is identical to the approximation in the FIFO case if we "eliminate" all make-to-order products. The approximation of make-to-order products, which is based on the results of Theorem 3, uses the same principles as Approximations 2 and 3. The workload process now lives on the entire non-negative half-line. Moreover, rather than the relative traffic intensity ρ_0 which may be greater or equal to one, we propose to use the *actual* traffic intensity $\tilde{\rho}_0$ which will be strictly less than one.

Numerical Examples

For the case of FIFO networks, the approximations proposed in this section slightly differ from those given in our previous works [13, 14]. For example, whereas [13] approximates $\tilde{\alpha}_1$ by

$$1 - \frac{1}{\lambda_1 m_{01}} \left(\frac{\rho_0 - 1}{1 - e^{-n_1 \theta / \tilde{\alpha}_1 \lambda_1}} \right),$$

our present approximation contains an extra ρ_0 in the exponent term:

$$1 - \frac{1}{\lambda_1 m_{01}} \left(\frac{\rho_0 - 1}{1 - e^{-n_1 \rho_0 \theta / \tilde{\alpha}_1 \lambda_1}} \right).$$

The modifications to the approximation formulas were motivated by our heavy traffic analysis of priority service. Recall that the formulas presented in Section 4 are asymptotically exact

in the heavy traffic limit, which requires $n \to \infty$ and $\rho_0 \to 1$. Away from the limit, one must restore the finite base-stock levels n_k and the utilization factor ρ_0 to their appropriate roles in order to obtain good approximations. For technical reasons that we will not detail here, we realized that the formulas are more "correct" when the factor ρ_0 is introduced in the exponent term. This modification also results in much improved estimates.

In addition, we proposed in this paper an explicit formula for the fill rate of non-bottleneck job types, whereas [14] established only bounds. To test these approximations, let us consider a network with two MTS job types and one MTO job type ($d = 2$ and $c = 3$). We will study two systems whose service time and inter-demand time distributions are all assumed to be exponential and whose parameters are shown in Table 1. (These are the same systems studied in [14].) Note that both Systems I and II satisfy the product-form conditions (see Section 2) so we can compare our estimates against exact performance measures.

The performance measures of interest are the throughput rate and FGI of make-to-stock jobs and waiting time for make-to-order jobs. Tables 2 and 3 compare our approximations of these performance measures against exact figures. In most cases, the approximations are reasonably good. The method performs less well when the system is far from the heavy traffic limit (for example, when base-stock levels are low) or when there is no clear bottleneck job type (i.e., when $\lambda_1^{-1} n_1$ and $\lambda_2^{-1} n_2$ are approximately equal). As one would expect, the estimates improve markedly as the utilization of the workstation increases, as the base-stock levels increase, or as a clear bottleneck job type emerges.

To investigate the accuracy of the approximations for priority policies, let us consider a network consisting of two job types, one make-to-stock and the other make-to-order (thus $d = 1$ and $c = 2$). We will work with two such systems, whose parameters are presented in Table 4, and for each system we consider three base stock levels, $n_1 = 10, 20, 50$. We set the utilization of the workstation to be 0.90 in both systems. All processing times and inter-demand times are exponentially distributed.

We compare our estimates against those obtained from simulation. Table 5 displays the results when priority is given to make-to-order jobs. Because of their higher priority, make-to-order jobs are essentially unaffected by the number of make-to-stock jobs in the network. (One can think of these jobs as experiencing an M/M/1 queue.) The long-run average delay time for make-to-order jobs was found to be 0.13 in System III and 1.63 in System IV for all base-stock levels. Our approximation method did extremely well in estimating these performance measures — the estimates were virtually exact — as one would expect from any reasonable approximation method, so we did not include these figures in Table 5.

The percentages shown in Table 5 have the following interpretations: the percentage as-

		Fill Rate: Type 1 jobs			Fill Rate:Type 2 jobs		
n_1	n_2	HT	Exact	% Err	HT	Exact	% Err
System 1							
5	10	85.4%	85.7%	-0.4%	97.9%	99.3%	-1.4%
10	20	93.1%	93.1%	0.0%	99.3%	100.0%	-0.7%
15	30	95.9%	95.9%	0.0%	99.7%	100.0%	-0.3%
16	30	96.2%	96.2%	0.0%	99.6%	100.0%	-0.4%
32	60	98.8%	98.8%	0.0%	99.9%	100.0%	-0.1%
48	90	99.5%	99.5%	0.0%	100.0%	100.0%	0.0%
16	20	96.2%	96.4%	-0.2%	98.8%	99.7%	-0.9%
32	40	98.8%	98.8%	0.0%	99.7%	100.0%	-0.3%
48	60	99.5%	99.5%	0.0%	99.9%	100.0%	-0.1%
10	10	93.1%	94.4%	-1.4%	96.2%	97.3%	-1.1%
20	20	97.3%	97.7%	-0.4%	98.5%	99.2%	-0.7%
30	30	98.7%	98.8%	-0.1%	99.3%	99.7%	-0.5%
System 2							
10	10	92.2%	95.2%	-3.1%	92.2%	95.2%	-3.1%
20	20	96.5%	97.9%	-1.4%	96.5%	97.9%	-1.4%
30	30	98.0%	98.8%	-0.8%	98.0%	98.8%	-0.8%
40	40	98.7%	99.3%	-0.5%	98.7%	99.3%	-0.5%
50	50	99.2%	99.5%	-0.4%	99.2%	99.5%	-0.4%
10	20	92.2%	92.4%	-0.2%	98.6%	99.8%	-1.1%
10	30	92.2%	92.2%	0.0%	99.6%	100.0%	-0.4%
10	40	92.2%	92.2%	0.0%	99.9%	100.0%	-0.1%
10	50	92.2%	92.2%	0.0%	100.0%	100.0%	0.0%
20	50	96.5%	96.5%	0.0%	99.7%	100.0%	-0.3%
30	50	98.0%	98.0%	0.0%	99.6%	100.0%	-0.4%
40	50	98.7%	98.9%	-0.1%	99.4%	99.8%	-0.4%

Table 2: Fill rate estimates for Systems I and II

		MTO Waiting Time			FGI at Station 1			FGI at Station 2		
n_1	n_2	HT	Exact	% Err	HT	Exact	% Err	HT	Exact	% Err
System 1										
5	10	0.34	0.33	2.6%	2.63	2.67	-1.6%	7.87	7.60	3.6%
10	20	0.59	0.59	0.7%	5.46	5.51	-0.8%	16.22	15.93	1.8%
15	30	.83	.83	0.0%	8.51	8.57	-0.7%	24.73	24.48	1.0%
16	30	.87	.87	0.4%	9.14	9.20	-0.6%	24.45	24.21	1.0%
32	60	1.44	1.44	0.6%	20.30	20.40	-0.5%	50.75	50.64	0.2%
48	90	1.82	1.81	0.6%	33.13	33.24	-0.3%	78.33	78.30	0.0%
16	20	0.87	0.85	2.2%	9.14	9.26	-1.3%	14.50	14.34	1.1%
32	40	1.44	1.44	0.6%	20.30	20.41	-0.5%	30.78	30.67	0.4%
48	60	1.82	1.81	0.6%	33.13	33.24	-0.3%	48.34	48.30	0.1%
10	10	0.59	0.54	10.0%	5.46	5.73	-4.6%	6.34	6.47	-2.0%
20	20	1.03	0.99	4.4%	11.75	12.02	-2.2%	13.47	13.54	-0.5%
30	30	1.38	1.36	2.1%	18.80	19.04	-1.2%	21.20	21.23	-0.2%
System 2										
10	10	0.51	0.43	19.0%	5.27	5.84	-9.6%	5.27	5.84	-9.6%
20	20	0.92	0.82	12.1%	11.04	11.84	-6.8%	11.04	11.84	-6.8%
30	30	1.28	1.18	9.1%	17.29	18.21	-5.1%	17.29	18.21	-5.1%
40	40	1.60	1.50	7.2%	23.99	24.97	-3.9%	23.99	24.97	-3.9%
50	50	1.88	1.78	5.9%	31.13	32.11	-3.0%	31.13	32.11	-3.0%
10	20	0.51	0.51	-1.4%	5.27	5.32	-0.9%	14.94	14.56	2.6%
10	30	0.51	0.52	-2.5%	5.27	5.30	-0.5%	24.90	24.47	1.8%
10	40	0.51	0.52	-2.5%	5.27	5.30	-0.5%	34.88	34.47	1.2%
10	50	0.51	0.52	-2.5%	5.27	5.30	-0.5%	44.88	44.47	0.9%
20	50	0.92	0.93	-1.1%	11.04	11.09	-0.4%	40.74	40.38	0.9%
30	50	1.28	1.29	-0.4%	17.29	17.36	-0.4%	37.09	36.78	0.8%
40	50	1.6	1.59	1.2%	23.99	24.24	-1.0%	33.88	33.88	0.0%

Table 3: Waiting time and inventory estimates for Systems I and II

System	Make-to-stock		Make-to-order		
	λ_1	m_{01}	λ_2	m_{02}	ρ_0
III	1.0	0.6	1.0	0.3	0.900
IV	1.0	0.2	1.0	0.7	0.900

Table 4: System Parameters: Systems III and IV

n_1	System III				System IV			
	Fill Rate		FGI		Fill Rate		FGI	
	Sim	Approx	Sim	Approx	Sim	Approx	Sim	Approx
10	0.949	0.950	6.31	6.19	0.758	0.532	5.46	5.79
	0.40%	0.11%	1.31%	1.90%	0.83%	29.82%	0.51%	6.04%
20	0.990	0.990	14.39	14.19	0.870	0.804	12.57	12.06
	0.62%	0.00%	1.02%	1.39%	0.83%	7.59%	0.75%	4.06%
50	1.000	1.000	42.68	42.90	0.965	0.962	37.00	34.93
	0.67%	0.00%	1.08%	0.52%	0.77%	0.31%	1.16%	5.59%

Table 5: Priority service for make-to-order jobs

sociated with a simulation number corresponds to the 95% confidence interval expressed as a percentage of the number, whereas the percentage associated with an approximation is the relative error of the estimate as compared with simulation. The estimates of long-run average fill rate and FGI all fall within the 95% confidence iterval for System III. The approximation does not perform as well in System IV but the estimates improve as the base stock level increases (i.e., as the system approaches heavy traffic).

Table 6 compares approximated and simulated queue times for make-to-order jobs when make-to-stock jobs receive processing priority. As in the previous scenario, our approximation performs extremely well in predicting throughput rate and FGI for make-to-stock jobs so we do not present those numbers here. (The throughput rates are essentially 1.00 for all base-stock levels; the base stock levels are essentially $n_1 - 1.5$ and $n_1 - .25$ for System III and System IV, respectively; and the estimates are exact up to 3 decimal places.)

5 Setting the Base-Stock Levels, Revisited

The problem of setting base-stock levels is considerably more complex when there are multiple make-to-stock job types because our characterization of fill rates is not as explicit. In particular,

	System III		System IV	
n_1	Sim	Approx	Sim	Approx
10	10.40	11.19	6.46	6.63
	6.75%	7.60%	4.91%	2.63%
20	11.20	11.25	6.66	6.63
	5.35%	0.45%	7.39%	0.45%
50	11.40	11.25	6.45	6.63
	6.73%	1.32%	6.50%	2.79%

Table 6: Waiting time under priority service for make-to-stock jobs

our estimates of fill rates depend not only on the values of the base stock levels but also on their *relative* magnitude, as measured by the ratios n_k/λ_k, where n_k and λ_k are the base-stock level and demand rate of type k jobs, respectively. These ratios determine "bottleneck" versus "non-bottleneck" job types, and our estimation method differs for a given job type depending on its bottleneck/non-bottleneck status. This leaves us with the unresolved issue of which type to designate as the bottleneck job type.

Denote by α_k be the specified fill rate for job type k, $k = 1, \ldots, d$. Define the ratios

$$r_k = \begin{cases} \dfrac{\sigma^2}{2\lambda_k m_{0k}} \left(\dfrac{\alpha_k}{1-\alpha_k} \right) & \text{if } \rho_0 = 1 \\ \left(\dfrac{\sigma^2}{2\rho_0(1-\rho_0)} \right) \alpha_k \ \ln \left(1 + \dfrac{1-\rho_0}{\lambda_k m_{0k}(1-\alpha_k)} \right) & \text{otherwise.} \end{cases} \tag{41}$$

Let us for now suppose that $d = 2$ and suppose that job types are numbered with $r_1 \leq r_2$. We propose the following rule-of-thumb in determining the bottleneck job type:

(a) If $r_1 < 0.8 r_2$ then type 1 is the bottleneck job type and type 2 is non-bottleneck.

(b) Otherwise, we consider r_1 to be "approximately equal" to r_2, in which case we think of both types as being non-bottleneck job types.

The base-stock levels are then determined as follows:

(a) If k is a bottleneck job type, then

$$N_k(\alpha_k) = \begin{cases} \left\lfloor \dfrac{\lambda_k \sigma^2}{2\lambda_k m_{0k}} \left(\dfrac{\alpha_k}{1-\alpha_k} \right) \right\rfloor + 1 & \text{if } \rho_0 = 1 \\ \left\lfloor \left(\dfrac{\lambda_k \sigma^2}{2\rho_0(1-\rho_0)} \right) \alpha_k \ \ln \left(1 + \dfrac{1-\rho_0}{\lambda_k m_{0k}(1-\alpha_k)} \right) \right\rfloor + 1 & \text{otherwise;} \end{cases} \tag{42}$$

$$N_k^r(\alpha_k) = \left\lfloor \frac{1}{1-\rho_r} N_k(\alpha_k) \right\rfloor + 1; \tag{43}$$

$$N_k^s(\alpha_k) = \left\lfloor \left(\frac{\lambda_k \sigma^2}{2\rho_s(1-\rho_s)} \right) \alpha_k \ln\left(1 + \frac{1-\rho_s}{\lambda_k m_{0k}(1-\alpha_k)} \right) \right\rfloor + 1. \tag{44}$$

(b) If k is a non-bottleneck job type, then

$$N_k(\alpha_k) = \left\lfloor \left(\frac{\lambda_k \sigma^2}{2\hat{\rho}_0^k(1-\hat{\rho}_0^k)} \right) \alpha_k \ln\left(1 + \frac{1-\hat{\rho}_0^k}{\lambda_k m_{0k}(1-\alpha_k)} \right) \right\rfloor + 1; \tag{45}$$

$$N_k^r(\alpha_k) = \left\lfloor \frac{1}{1-\rho_r} N_k(\alpha_k) \right\rfloor + 1; \tag{46}$$

$$N_k^s(\alpha_k) = \left\lfloor \left(\frac{\lambda_k \sigma^2}{2\hat{\rho}_s(1-\hat{\rho}_s)} \right) \alpha_k \ln\left(1 + \frac{1-\hat{\rho}_s}{\lambda_k m_{0k}(1-\alpha_k)} \right) \right\rfloor + 1, \tag{47}$$

where

$$\hat{\rho}_0^k \equiv \sum_{l=1}^{k} \alpha_l \lambda_l m_{0l} + \sum_{l=k+1}^{c} \lambda_l m_{0l} \quad \text{and} \quad \hat{\rho}_s^k \equiv \sum_{l=1}^{k} \alpha_l \lambda_l m_{0l} + \sum_{l=k+1}^{d} \lambda_l m_{0l}. \tag{48}$$

Given the discussions in the previous sections, it is straightforward to extend these approximations to many ($d > 2$) make-to-stock job type. Let us therefore devote our attention to the implications of the base-stock estimates.

Table 7 displays the base-stock levels required to achieve various service levels for the two systems described in Table 1. The desired service levels are given in the first column. The second column contains the computed base-stock levels for each job type, where N_k^b (respectively, N_k^n) is the base-stock level for type k jobs if type k jobs were viewed as the bottleneck (respectively, non-bottleneck) job type. The third column shows the computed ratios r_k, and the fourth column contains the recommended base-stock levels using the decision criteria outlined at the beginning at the section. The exact base-stock levels required to achieve the specified service levels were determined by exhaustive search (recall that these systems have product-form solutions) and are shown in the last column.

The designation of bottleneck versus non-bottleneck job type can have significant impact on the recommended base-stock level. If all jobs were treated as bottleneck job types, the resulting approximations would require much higher base-stock levels than necessary. On the other hand, if all jobs were treated as non-bottleneck job types, the resulting recommended base-stock levels would not be enough to deliver the required service level.

Moreover, the numbers in Table 7 illustrate the conventional wisdom that it is typically very costly to require high service levels for *all* make-to-stock job types. To be specific, let

α_1, α_2	N_1^b	N_1^n	N_2^b	N_2^n	r_1, r_2	N_1, N_2	N_1^*, N_2^*
System I							
90%,90%	8	6	7	5	0.91,1.10	6,5	6,5
90%,95%	8	6	13	8	0.91,1.93	8,8	7,7
90%,99%	8	6	31	16	0.91,4.84	8,16	7,11
95%,90%	14	9	7	5	1.63,1.10	9,7	9,6
95%,95%	14	10	13	9	1.63,1.93	10,9	10,9
95%,99%	14	11	31	20	1.63,4.84	14,20	12,14
99%,90%	35	20	7	6	4.34,1.10	20,7	14,7
99%,95%	35	25	13	10	4.34,1.93	25,13	19,11
99%,99%	35	30	31	26	4.34,4.84	30,26	28,23
System II							
90%,90%	8	6	8	6	0.79,0.79	6,6	6,6
90%,95%	8	6	15	9	0.79,1.49	8,9	6,9
90%,99%	8	6	46	20	0.79,4.55	8,20	8,14
95%,95%	15	10	15	10	1.49,1.49	10,10	10,10
95%,99%	15	12	46	26	1.49,4.55	15,26	14,20
99%,99%	46	36	46	36	4.55,4.55	36,36	35,35

Table 7: Base-stock levels

us focus on System II. The exact base-stocks required to achieve service levels of $(90\%, 99\%)$, $(95\%, 99\%)$ and $(99\%, 99\%)$ are $(8, 14)$, $(14, 20)$, $(35, 35)$, respectively. (The recommended base-stock levels are similar.) Note that a higher service level for type 1 jobs translates into not only a higher type 1 base-stock but also higher base-stock for type 2. In this example, the base-stock level for type 2 jobs experiences a more than two-fold increase when type 1 service level is changed from 90% to 99%.

From the examples presented in this paper, it is evident that the problem of how to set base-stock levels in a hybrid make-to-stock/make-to-order production environment involves complex and subtle issues. This work provides simple and reliable estimates of performance measures for such systems, allowing managers to quickly evaluate the impact of different operating scenarios and understand the trade-offs between service and operating costs.

A Proof of Theorem 2

We will present here an outline of the proofs for Theorems 2 and 3. The details of the proofs are typical of heavy traffic analyses so in the interest of brevity, we will suppress them whenever possible and refer readers to previous works for the full proofs. For the purposes of the proofs, it will be convenient to assume that initially, the finished-goods-inventories are full and the workstation is empty.

Because the workstation (station 0) awards preemptive priority to open jobs, these jobs are unaffected by the presence of closed jobs and experience the queue in light traffic. Lemma 2 of Peterson [15] applies directly to prove that $Q_{0k}^n \Longrightarrow \zeta$ for $k = d+1, \ldots, c$, where $\zeta(t) = 0$ for all $t \geq 0$. Equivalently, denoting by $W_r(t)$ the amount of make-to-order high priority work at station 0 at time t, we have $W_r^n \Longrightarrow \zeta$. We now follow the same approach by Peterson [15] to prove convergence for the processes of lower priority.

For $k = 1, \ldots, c$, denote by $V_{0k}(i)$ the sum of the first i type k service times at the workstation and by $A_k(t)$ the number of class k arrivals to the workstation by time t. For $k = 1, \ldots, d$, let $V_k(i)$ be the sum of the first i service times at station k following the *initial* n_k jobs that are at station k at time 0. Set

$$E(t) \equiv \sup_{0 \leq s \leq t} \left[\sum_{k=d+1}^{c} V_{0k}(A_k(s)) - s \right]^-,$$

interpreting $E(t)$ as the amount of time available for processing low-priority make-to-stock products in the first t units of time. Letting

$$\hat{E}(t) \equiv \hat{E}(t) - \rho_s t, \quad \textit{and}$$

$$\hat{E}^n(t) \equiv \frac{1}{n}\hat{E}^{(n)}(n^2 t),$$

it follows from Peterson [15] that $\hat{E}^n \Longrightarrow \hat{E}^*$ where $\hat{E}^*$ is Brownian motion with drift μ (from equation (10)) and variance σ^2 (from equation (5)).

For $k = 1, \ldots, d$, letS_k be the renewal process associated service times at station k; equivalently, S_k represents the inter-demand process of (make-to-stock) type k jobs. For $k = 1, \ldots, c$, let S_{0k} be the renewal process associated with type k service times at the workstation; for $j = 0, \ldots, d$, let $B_j(t)$ be the amount of time station j is busy in $[0, t]$; and let $I_j(t) \equiv t - B_j(t)$ be the cumulative idleness process for station j. If we let $T_k(t)$ be the amount of time the workstation has devoted to type k service during the first t units of time, it follows from the previous definitions that

$$\sum_{k=1}^{d} T_k(t) = E(t) - I_0(t).$$

For this section, let us denote by W_k the workload process at station k ($k = 0, \ldots, d$). The amount of work associated with the initial n_k jobs at station k ($k = 1, \ldots, d$) is accordingly denoted as $W_k(0)$. We have

$$W_k(t) = \begin{cases} \sum_{k=1}^{d} V_{0k}(S_k(B_k(t))) + \sum_{k=d+1}^{c} V_{0k}(A_k(t)) - B_0(t) & \text{if } k = 0, \\ W_k(0) + V_k(S_{0k}(T_k(t))) - B_k(t) & \text{if } k = 1, \ldots, d. \end{cases} \tag{49}$$

Let us now define the "centered" processes

$$\begin{aligned} \hat{V}_{0k}(t) &\equiv V_{0k}(t) - m_{0k}t & \hat{S}_{0k}(t) &\equiv S_{0k}(t) - (1/m_{0k})t, \\ \hat{V}_k(t) &\equiv V_k(t) - (1/\lambda_k)t & \hat{S}_k(t) &\equiv S_k(t) - \lambda_k t, \\ \hat{A}_k(t) &\equiv A_k(t) - \lambda_k t & \hat{T}_k(t) &\equiv T_k(t) - \lambda_k m_{0k} t. \end{aligned}$$

Define $\eta(t)$ to be the arrival time to the workstation of the low priority job receiving service at station 0 at time t, and set $\eta(t) \equiv t$ if no low priority job is in service at that time. The amount of low-priority time available between $\eta(t)$ and t is exactly the amount of low-priority work present at time $\eta(t)$ less the remaining service time of the low priority job currently in service, if there is one. That is,

$$E(t) - E(\eta(t)) = [W_0(\eta(t)) - W_r(\eta(t))] - \epsilon_1(t),$$

where $\epsilon_1(t)$ represents the remaining service time if the job currently in service is low-priority, and is zero otherwise. In terms of centered processes, we have

$$\rho_s(t - \eta(t)) = \left[\hat{E}(\eta(t)) - \hat{E}(t)\right] + [W_0(\eta(t)) - W_r(\eta(t))] - \epsilon_1(t).$$

Similarly, the allocation processes obey the property

$$T_k(t) = V_{0k}(S_k(B_k(\eta(t)))) + \epsilon_{2k}(t), \quad 1 \le k \le d, \tag{50}$$

where $\epsilon_{2k}(t)$ is the amount of service the current job has received if that job is of type k and is zero otherwise. Writing (50) in terms of centered processes, we arrive at

$$\begin{aligned} \hat{T}_k(t) &= \hat{V}_{0k}(S_k(B_k(\eta(t)))) + m_{0k}\hat{S}_k(B_k(\eta(t))) - \lambda_k m_{0k} I_k(\eta(t)) - \\ &\quad \frac{\lambda_k m_{0k}}{\rho_s}\left\{\hat{E}(\eta(t)) - \hat{E}(t) + W_0(\eta(t)) - W_r(\eta(t)) - \epsilon_1(t)\right\} + \epsilon_{2k}(t). \end{aligned} \tag{51}$$

Moreover, (49) can be written as

$$W_k(t) = \begin{cases} \sum_{k=1}^{d}\left[\hat{V}_{0k}(S_k(B_k(t))) + m_{0k}\hat{S}_k(B_k(t)) - \lambda_k m_{0k} I_k(t)\right] + \\ \sum_{k=d+1}^{c}\left[\hat{V}_{0k}(A_k(t)) + m_{0k}\hat{A}_k(t)\right] + n(\rho_0 - 1)t + I_0(t) & \text{if } k = 0, \\ W_k(0) + \hat{V}_k(S_{0k}(T_k(t))) + \frac{1}{\lambda_k}\hat{S}_{0k}(T_k(t)) + \\ \frac{1}{\lambda_k m_{0k}}\hat{T}_k(t) + I_k(t) & \text{if } k = 1, \ldots, d. \end{cases} \tag{52}$$

Let us introduce the following scaling conventions: for $X^{(n)}$ a generic process, then

$$X^n(t) \equiv \frac{1}{n}X^{(n)}(n^2 t) \qquad \bar{X}^n(t) \equiv \frac{1}{n^2}X^{(n)}(n^2 t).$$

Setting

$$\xi_0(t) \equiv \begin{cases} \sum_{k=1}^{d}\left[\hat{V}_{0k}(S_k(B_k(t))) + m_{0k}\hat{S}_k(B_k(t))\right] + \\ \sum_{k=d+1}^{c}\left[\hat{V}_{0k}(A_k(t)) + m_{0k}\hat{A}_k(t)\right] + n(\rho_0 - 1)t & \text{if } k = 0, \\ \hat{V}_k(S_{0k}(T_k(t))) + \frac{1}{\lambda_k}\hat{S}_{0k}(T_k(t)) + \\ \frac{1}{\lambda_k m_{0k}}\left\{\hat{V}_{0k}(S_k(B_k(\eta(t)))) + m_{0k}\hat{S}_k(B_k(\eta(t)))\right\} - \\ \frac{1}{\rho_s}\left\{\hat{E}(\eta(t)) - \hat{E}(t) + \xi_0^n(\eta(t)) - W_r(\eta(t)) - \epsilon_1(t)\right\} + \\ \frac{1}{\lambda_k m_{0k}}\epsilon_{2k}(t) & \text{if } k = 1, \ldots, d, \end{cases}$$

the scaled workload processes can be written as

$$W_0^n(t) = \xi_0^n(t) + I_0^n(t) - \sum_{k=1}^{d}\lambda_k^{(n)} m_{0k}^{(n)} I_k^n(t); \tag{53}$$

$$W_k^n(t) = \xi_k^n(t) - \frac{1}{\rho_s^{(n)}}\left(I_0^n(t) + \sum_{k=1}^{d}\lambda_k^{(n)} m_{0k}^{(n)} I_k^n(t)\right) + \left[I_k^n(t) - I_k^n(\bar{\eta}^n(t))\right], \quad k = 1, \ldots, d; \tag{54}$$

$$I_j^n(\cdot) \text{ is continuous an nondecreasing with } I_j^n(0) = 0,\ j = 0, \ldots, d; \tag{55}$$

$$I_j^n(\cdot) \text{ increases only at times } t \text{ when } W_j^n(t) = 0,\ j = 0, \ldots, d. \tag{56}$$

The convergence of W^n to the desired limits follows from the results of Nguyen [14]. The convergence of scaled queue length processes Q_{0k}^n and Q_k^n are proved similary with the methodology of Peterson [15]. □

B Proof of Theorem 3

Closed jobs receiving preemptive priority at station 0 are not affected by open jobs, so the dynamics of closed jobs in this system may be described a a closed network with type k jobs circulating between station 0 and station k. The results from Nguyen [14] can therefore be applied directly to obtain convergence for the scaled queue length processes Q^n_{0k} and Q^n_k, $k = 1, \ldots, d$. Turning to the total workload at station 0 and using the notation developed in Appendix A, we have

$$W^n_0(t) = \xi^n_0(t) + I^n_0(t) - \sum_{k=1}^{d} \lambda_k m_{0k} I^n_k(t).$$

The results from analysis of high priority make-to-stock jobs imply $I^n_k \Longrightarrow \zeta$ where $\zeta(t) = 0$ for all $k = 1, \ldots, d$ and $t \geq 0$. We can therefore can conclude that W^n converges to an RBM on the half line $[0, \infty)$ with drift μ and variance σ^2. Convergence for the queue length of make-to-order jobs follow from the methods of Peterson [15]. □

References

[1] Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* **22**, 248–260 (1975).

[2] Billingsley, P. *Convergence of Probability Measures.* Wiley, New York, 1968.

[3] Buzacott, J.A. and Shanthikumar, J.G. *Stochastic Models of Manufacturing Systems.* Prentice Hall, Englewood Cliffs, 1993.

[4] Carr, S. A., Güllü, A. R., Jackson, P. R. and Muckstadt, J. *An Exact Analysis of the No B/C Stock Policy,* preprint, 1993.

[5] Chen, H. and Mandelbaum, A. Stochastic discrete flow networks: Diffusion approximation and bottlenecks. *Annals of Probability* **19**, 1463–1519 (1991).

[6] Dai, J. G. and Harrison J. M. The QNET method for two-moment analysis of closed manufacturing systems. Submitted for publication (1992).

[7] Glynn, P. W. Diffusion Approximations. In *Handbook on OR and MS,* Vol 2., D. P. Heyman and M. J. Sobel (eds.), Elsevier Science, North Holland, 145–198 (1990).

[8] Harrison, J. M. *Brownian motion and stochastic flow systems.* Wiley, New York, 1985.

[9] Harrison, J. M. and Nguyen, V. The QNET method for two-moment analysis of open queueing networks. *Queueing Systems: Theory and Applications* **6**, 1-32 (1990).

[10] Harrison, J. M. and Nguyen, V. Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems: Theory and Applications* (forthcoming).

[11] Iglehart, D. L. and Whitt W. Multiple channel queues in heavy traffic I. *Advances in Applied Probability* **2**, 150–177 (1970).

[12] Kelly, F. P. *Reversibility and Stochastic Networks.* Wiley, New York, 1979.

[13] Nguyen, V. Fluid and Diffusion Approximations of A Two-Station Mixed Queueing Network. *Mathematics of Operations Research*, to appear.

[14] Nguyen, V. A Multiclass Hybrid Production Center in Heavy Traffic. Submitted for publication.

[15] Peterson, W. P. A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research* **16**, 90–118 (1991).

[16] Whitt, W. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, **5** 67–85 (1980).

[17] Yao, D.D. Ed. *Stochastic Modeling and Analysis of Manufacturing Systems*. Springer-Verlag, New York, 1994.